Research Paper | Education | E-ISSN No : 2454-9916 | Volume : 9 | Issue : 10 | October 2023

CYBERSECURITY: A MACHINE LEARNING APPROACH

Prof. D. S. Chandak[1], Mr. Rajat Dubey[2]

[1]Assistant Professor,PRMCEAM-Badnera-Amravati,
[2] Allianz Commercial Austin-USA

ABSTRACT

Machine Learning is one of the effective responses to trivial attacks, starting with the Internet Protocol traffic classification and the filtering of mistreatment traffic for intrusion detection. The latest study is being conducted out using traffic data and Machine Learning techniques. This article presents a study of the literature on machine learning and its uses in online-based data security frameworks for malware prevention, labeling, and utilization like email filtering. Each approach has been defined and synthesized based on the significance and amount of citation. Due to the importance of datasets in the Machine Learning methods, many well-known databases are often referenced. There are also some guidelines for using a particular algorithm. The MODBUS data obtained from a gas pipeline were evaluated using four Machine Learning algorithms. Several attacks were categorized by Machine Learning algorithms and each algorithm was ultimately judged for its efficiency.

KEYWORDS: Cyber Security, Dataset, Data Sorting, Machine Learning Algorithms.

1. INTRODUCTION

This article is a literary review of computer security systems' machine learning and data mining approaches. Few ML techniques are defined in conjunction with their cybersecurity use. The paper provides a number of comparative guidelines for the ML approach and suggestions on the best possible method were made according to the characteristics of cybersecurity problems. Second, the efficacy of five distinct algorithms when applied to ICS networks was compared using a MODBUS data collection. The Receiver Operating Characteristics (ROC) curve is often used to opt- out and discard optimum models regardless of cost content or class distribution. A ROC curve has therefore been developed to evaluate the accuracy of the binary classifier used for the under analysis data collection [1].

This study is intended for scholars who are interested in machine learning and computer security. Along with the concept of machine learning, several noteworthy works have been discussed, as well as some helpful explanations of how ML is often used to solve cyber problems. Since the early 2000s, a number of significant surveys on machine learning have been reported. Nguyen et al. investigate Network data classification techniques that aren't focused mostly on internet protocols or packet payloads in depth. This paper looks at how to classify IP addresses using machine learning and mathematical traffic characteristics [2]. Sperotto et al. conduct a comprehensive analysis of all major e- mail scanning and machine learning tools that might be used to distinguish and distinguish spam email from legitimate emails. The jurisdiction writings on such attacks have been outlined, and a detailed comparison of all of the approaches has been carried out [3]. For network intrusion detection, Tedero et al. propose computational, deep learning, and information frameworks. It emphasizes mostlyoutlier detection rather than biometrics detection [4]. Almomani et. al. utilized data from NetFlow and demonstrated, when the volume of network traffic exceeds any ceiling, the packet processing might be impossible at the rate of streaming. There are significant works that summarize the most recent research on machine learning and its use in the field of cybersecurity [5].

This paper is dedicated to understanding virtual safety datasets which can be utilized by the scientists interested in calculations that may be carried out to virtual express issues. A collection of machine learning algorithms was analysed for different attacks when testing the Remote Terminal Unit (RTU) in a proposed pipeline in the later part of the document on the gathered ICS data set. The data collection used consists of 35 distinct types of simulated ICS strikes. The consistency of each ML algorithm was tested in the divergence of anomalous attacks.

2. TAXONOMY OF THE RESEARCH

In ML methods, data is extremely important and a researcher needs to consider the data before doing any study. Secondly, in ML analytics it is not direct to use unprocessed data as packet capture (pcap), various network-based data and NetFlow. The data must be prepared for use in common ML tools such as WEKA, RapidMiner and R. As a result, researchers using ML framework analysis will think of the methods and approaches used in the data before analysing it. This article describes a low-level database and a traditional example for collecting network data [6]. Web protocols are used by applications operating on the level of users (144 according to the Internet Engineering Task Force). The primary mode of communication is data packets. Network traffic can be collected and stored in the packet collection form, received and distributed via the interfaces (physical and WLAN). Libpcap and Wincap are very common network utilities for UNIX and Windows. Some tools, such as tcpdump and Wireshark, can be used as a sniffer, protocol analyser, and network controller [7]. There are different properties and attributes of the data collection for deep learning. These attributes determine the primary features of each dataset. Therefore, when a lot of unprocessed data is collected by the scholar as a pcap, a script needs to be written to separate the required attributes from the pcap into the format of the ML method. Fowler et. al. have been investigating Weka's arff (Attribute- Relation File Format) and have created a method for converting any pdml (Packet Details Markup Language) format to arff format. To extrude a pcap report to pdml apparatuses such

as T-shark can be applied [8].

T-shark to T-pdml – R <inputs> <final outcome> ....(1)

Here the info document is the pcap record and the name of the pdf document is the yield document. Furthermore, "pdml2arff.py" (accessible in GitHub) a Fowler's instrument, perhaps utilized to play out the last transformation.

pdml to arff.py <inputs> ....(2)

The pdml index is known as input file generating an arff index known as <Input record>. arff

The findings of Fowler demonstrate that the tool works well with standard TCP traffic and transforms the raw data into a functional weka format. In final research, with the MODBUS Protocol, the method was used to transform pcap files collected by the gas pipeline and all attributes were translated into string nominal attributes that were readable by WEKA but could not be analyzed in any way [8]. The network interface is monitored by CISCO, and the IP network traffic is collected while entering and exiting the interface by Cisco. Through reviewing the data generated, a network administrator may evaluate information like the source and destination traffic and service class.

There are three main elements in a traditional NetFlow architecture Flow Exporter- capture the flow-to-flow collectors and export the network traffic. The data is collected and pre-processed and then stored by the data collector, Application for Analysis – Split the flowing data and profile it as needed. The simplified and formatted version of current network packets is used in NetFlow info [9]. Some datasets in possession of the Defence Advanced Research Project Agency (DARPA) might be vital to the community protection scientists. Data sets from DARPA 1998 and DARPA 1999 were developed with the assistance of Cyber Systems and Technology gathered from the Lincoln Library of the Massachusetts Institute of Technology (MIT). Knowledge Discovery is also an informational index that is transcendently utilized in cybersecurity [10].

Another conspicuous informational index includes the SCADA conference produced with the aid of using the Mississippi State University's infrastructure coverage focus [1]. The particular data set can be tested within the later regions to assess the exactness of Machine Learning calculations at SCADA conventions. The informational series obtains the data from a flow of processed network pipeline and records thirty-five specific attacks at the SCADA framework.

2.1 ML Algorithms for Cyber Security

This paper discusses a small database and a typical example of network data collection.

2.1.1 Bayesian Network

The system is established into a directed acyclic graph as an altered variable set and its conditional dependencies. The parental nodes depend on the children's nodes, and each seed preserves the conditional probability form and random variable. Figure 1 demonstrates a Bayesian network attack signature identification. Each state is input with different status values to the underlying state [11]. The likelihood figures measured will be shown and calculated in the diagram. In addition to random heterogeneity, Bayesian networks can be used.

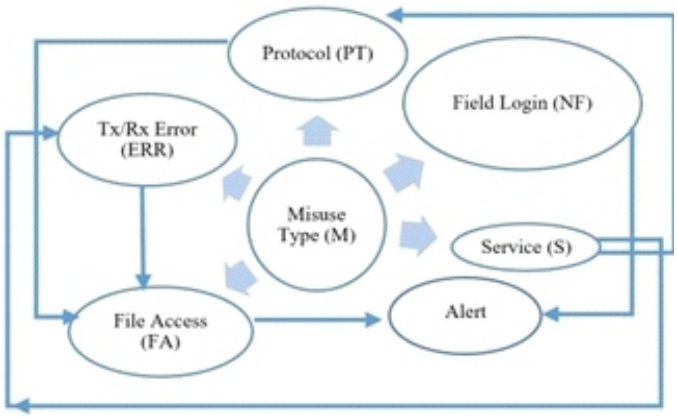


Figure 1. Bayesian networks are used to track attack signatures [1]

For identification of abnormality, the Bayesian network and the known threat pattern can also compare correlations with data processing for known attacks. Goyal et. al. has been developing a Bayesian network intrusion detection method.

File Access State input variables and values	P (FA = True)	P (FA = False)
M = R2H, PT = NSF, EPR = 0	0.95	0.05
M = R2H, PT = FTP, EPR = 0	0.99	0.01
M = Probe, PT = None, EPR = 50%	0.80	0.20
M = Probe, PT = PING, EPR = 0	0.50	0.50
M = DoS, PT = POP, EPR = 100%	0.80	0.20
M = DoS, PT = HTTP, EPR = 50%	0.90	0.10

Table 1. Detection Rate

The KDD 1999 was used for modelling the structure with 9 of its attributes. In standard scenarios and attack scenarios, 88% and 89% were reached. The detection rate for Samples, DOS and R2L was 99%, 21%, 89%, with the detection rate of 7%. The accuracy of the model was significantly impaired as the number of training instances for the R2L is very low [13].

2.1.2. Decision Trees

A tree is a good analogy for the decision tree. The leaves on the trees represent the different classifications, while the limbs represent the connections or characteristics that lead to the classifications. A few common algorithms for automatically creating decision trees are ID3 and C4.5. Because of a vast number of signatures, matching the SNORT laws with the incoming traffic takes a long time. Kruegel and Toth et al. used a version of the ID3 algorithm to replace 150 SNORT rules. They aimed to use a decision tree model to replace these algorithms. This would help raise processing speed. Snort rules were replaced with law clustering. This reduces the number of comparisons needed. Parallel assessment is also possible, which speeds up the comparative process. The install model was compared to the snort study in terms of processing speed and performance. The model hit a top speed of 105 percent and a bottom speed of 5%. To allow for further innovation, the set of actions substituted was increased from 149 to 1580+. About the fact that Toth does not have any objective statistics, the analysis discovered a significant speedup while using the decision tree system, as well as a significant reduction in processing time [15].

2.1.3. Clustering

Clustering algorithms may learn from audit results,so a system administrator's specific definition of various attack groups isn't needed. Hendry et al. illustrate the use of a clustering algorithm to detect real-time signatures. The Simple Logfile Clustering Tool (SLCT) has been utilized to establish natural and

anomalous internet traffic using a density-based clustering scheme. There are two clustering mechanisms in use: First, one scheme can be used to diagnose usual and attack conditions, and the other might be utilized to evaluate internet traffic in a controlled manner. In this model, parameter M describes the function found in the cluster. With the M parameter set to 96%, 97% of attack data is observed with a FAR of 15%[16].

2.1.4. Artificial Neural Network (ANN)

The ANN functions primarily in the same way as the human brain. A layer structure is used in the neural network. The neuron in the second layer of the network is actuated by the data input. This, in particular, outputs to the hierarchy's next layer. This continues until the final layer of the network produces the output. Hidden layers are a form of an internal network that is shielded from the outside world and plays animportant role in neural networks. Owing to the existence of local minima, one of the main drawbacks of neural networks is the lengthy learning time. This method was popular in the mid- 1990s, but as support vector machines (SVMs) became more popular, ANN began to fade. The importance of neural networks has risen again since the advent of convolutional neural networks [18]. Canady defines an ANN model that detects deviations using a multi-category classifier.

The data was generated using the Real Secure network monitor. The system had attack signatures built-in. Of the 10,000 confirmed assaults, about 3000 programs like the Satan and Internet Scanner have been replicated. All data pre-processing was performed using ICMP formats, ICMP formats, address of the source, target identify, authentication code, origin route, target port, original data frequency and type of information. The ANN was then trained using the standard and attack data from the analysis. During preparation and simulation conditions, Canady et al. record an error rate of 0.059 and 0.071, respectively. As a result, an RMS of 0.071 corresponds to a research step precision of 93%. The data is divided into two categories: natural and malicious traffic [17].

2.1.5. Genetic Analysis

GP (Genetic Programming) and GC (Genetic Computation) are two of the most common computing methods focused on the survival of the fittest theorem. These algorithms work for chromosome samples that evolve depending on the three fundamental operators - collection, intersect, and alteration. The algorithm starts with a random population and determines each individual's fitness value. This indicates each individual's ability to solve the current dilemma, and those with a higher likelihood have a better chance of being included in the gene pool. The next step, crossover, will be performed by two competent individuals, and then both will be subjected to mutation [18]. The chromosome with the best match of the two mutated individuals would bepassed on to the next generation.

Abhram et al. developed a classifier for attacks using a basic GP model. Gene Expression Programming (GEP), Linear Genetic Programming (LGP) and Multi Expression Programming (MEP) were three common GP models used in this study. As function sets, the model incorporated various mathematical operators. There were 24 different attack scenarios in the dataset, including four different kinds of attacks (U2R, R2L, DoS, and probing). Based on the kind of ambush being investigated, the above model's false alarm rate (FAR) ranged from 0% to 4.9% [19].

2.1.6. Hidden Markov Models

The topology of the model is determined by a series of conditions that are correlated by using transformation prospects. The secret parameters can be determined from the measurable parameters using the forward-backwards correlation provided by this model. Since each state's probability distribution differs, the system can shift states over time and can represent non-stationary sequences. HMM was used by Joshi et al. to establish an intrusion detection method. The interlinking between states has been structured in such a way that every state will transform into another. The Baum- Welch approach can be used to approximate HMM parameters. The KDD 1999 dataset was used to test the model's validity. Five attributes of the datasets were selected for review out of a total of 41 [20]. The model's identification rate was 78.9%, with a false positive rate of 20%. The accuracy of the model could increase if more than 5 features are included in the study.

2.1.7. Inductive Learning

The deduction is the process of inferring specific facts from a dataset. Inductive learning, on the other hand, is a method of going from basic perception to the creation of hypotheses and patterns. These are the two most common methods for extracting information from files. Inductive analysis generates some broad patterns that are then used to generate some hypothetical conclusions. To induce random events and anomalous traffic, artificial anomaly generation has been developed by Das et al. [1]. The model had an efficient 95 percent identification rate and a low false 2 percent warning rate. This assessment diagnosed the proper method to accumulate the dataset that may be applied for abnormality discovery and confirmed the usage of inductive gaining knowledge of version at the created dataset. This research outlines the right data- set approaches that can be used for the identification of anomalies and have shown the application of an inductive data-set learning model.

2.1.8. Outlier Detection

In cyber-security, machine learning is used in three major areas: intrusion detection systems, outlier detection systems, and malicious behaviour. Outlier detection distinguishes irregular traffic from regular traffic, while malware analysis classifies attack signatures by contrasting them to existing ones. A clustering algorithm (such as DBSCAN's density- based clustering algorithm) performs well for outlierdetection. Apart from the high frequency of processing, clustering algorithms are easy to apply and have fewer constraints to modify.

SVM is also very effective at detecting anomalies. The classifiers must be able to produce identities in order to detect misuse. Marks capable of such tasks are generated by branch focus in a choice tree or chromosomes in genetic estimation. As a consequence, no longer valid are formulas such as ANN and SVM, which involve updating nodes [18].

3. Discussion

The major purpose of the assessment is to see whether machine ML algorithms can be used to identify cyber-attacks on MODBUS info. The ML models were created by tenfold cross-validation and Weka [6] was used to do this experiment. Weka generates 10 various frameworks for the given database using 10 fold cross- validation. The end product is a weighted average of these models. The data were indeed named telemetry data from the gas pipeline established by the Digital Communication Protection Centre, Mississippi State University [1]. For the test, only a few common classifiers were used. The strategies applied were Naïve Bayes, Random Forest, OneR and J48.

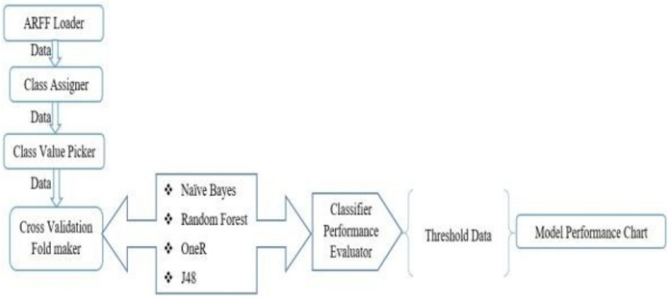
3.1. Information about the Dataset

Weka minable arff format was used for the dataset. It had a total

of 20 features and which has been shown in Table 2 lists all of the features found in the dataset.

Features	
CRC rate	specific result
control scheme	command response
address	reset rate
pump	Time
pressure measurement	categorized result
length	cycle time
solenoid	binary result
setpoint	Rate
gain	system mode
function	Deadband
length	cycle time

Table 2. Selected key features



Morris et al. provide an overview of every database layout and why most of these aspects relate to cyber protection. A total of 35 attacks were carried out in this dataset. Nave Malicious Response Injection (NMRI), Malicious Parameter Command Injection (MPCI), Complex Malicious Response Injection (CMRI), Malicious State Command Injection (MSCI), Denial of Service (DoS), Malicious Function Code Injection (MFCI) and Reconnaissance are the seven types of attacks.Theseseven attack categories, as well as standard traffic data, are included in the arff data set's final class. In the sample, there were 97019 instances. Table 3 below shows the representation of the last class.

Label of Class	Tabulate
Normal	61156
CMRI	15466
MPCI	7637
Reconnaissance	6805
NMRI	2763
DoS	1837
MSCI	782
MFCI	573

Table 3. Distribution of the final class

Receiver Operating Characteristics (ROC)

Beaver et al. have also used the ICS repository for ML analysis [21]. However, for every algorithm, no ROC curve was drawn, making it impossible to calculate the overall efficiency of the algorithms. The x-axis diagram of the false-positive incidence (FAR) versus the y-axis plot of test sensitivity is the receiver operating characteristics (ROC) curve. The ROC's field under the curve is a crucial framework. It helps to determine the sensitivity and specificity of a system. The true positive decisions number is known as the sensitivity, and the true negative decisions number is known as the precision.

Since the territory surrounding the ROC curve is an indicator of any test's overall success, this variable can be utilized to evaluate the effectiveness of the ML algorithms of MODBUS data classification. Therefore, an AUC evaluation of the ROC curve against various Machine-learning algorithms will expose the classification accuracy of the algorithms. For this study, the Weka Knowledge Flow (WKF) framework was designed. Figure 2 depicts the predicament. Figure 5 illustrates the ROC curves created by the four ML algorithms. The ROC curve reveals that the j48 algorithm produces the best classification results for the power system dataset. The AUC, Precision, and Recall of the four algorithms are demonstrated in table 4.

Figure 2. WKF Model for designing the receiver operating characteristics curves

First of all, this research was carried out as a binary classification task. Multiclass grouping is a different method that can be used. Second, the weighting factor of all 8 classes is taken into account in the equation containing the AUC, Recall and Precision seen in table 4. Every class should be examined separately, providing details about the framework's ability to distinguish every attack from one another and regular traffic. The findings in Table 4 demonstrate the model's ability to distinguish all traffic.

Algorithms	AUC	Precision	Recall
J48	0.97	0.972	0.972
NaiveBayes	0.956	0.939	0.944
OneR	0.877	0.872	0.874
RandomForest	0.983	0.985	0.984

Table 4. AUC, Precision and Recall

Figure 3 shows that the J48 does well in overall classification since the ROC curves' area under the curve value is closer to 1. Figure 5 in the appendix depicts the ROC graph created by Weka. In industrial control systems (ICS), the implementation efficiency of a machine learning intrusion detection system is important. As a result, the planning process must be streamlined even though newly streamed information is being trained promptly whereas the model tracks information in real-time. Therefore, the optimal precision and planning time pair to every context of the ICS is often suggested when choosing an algorithm.

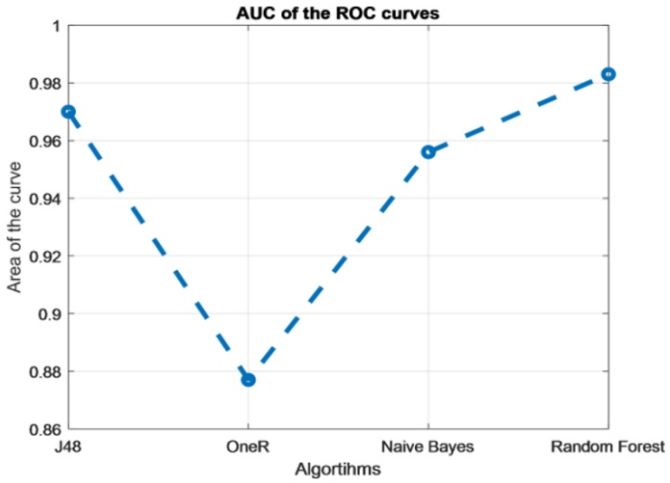


Figure 3. AUC curve for each of the four ML algorithms

A python script was employed to track the computation period of training during the K-fold validation. The calculated training time is the time it takes to verify all layers; therefore, in a realistic scenario, the processing time can be calculated as 1/Kth of the millisecond plotted time, as seen in Figure 4.

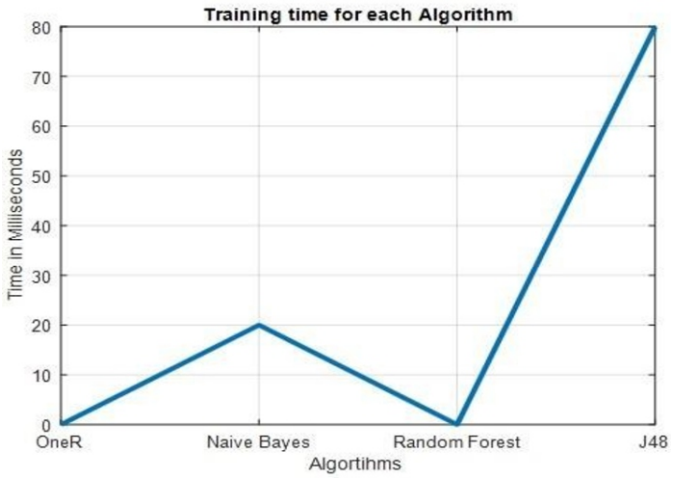


Figure 4. Each algorithm uses a different time to train

As a result, while J48 will outperform Random forest in terms of interpretability, when the schemes are being used as the basis of an interruption prevention framework a slight sacrifice in precision yields better real-time performance. . The training computer has an Intel i7 8550U main processor and Nvidia GeForce GTX 1050 dedicated graphics card.

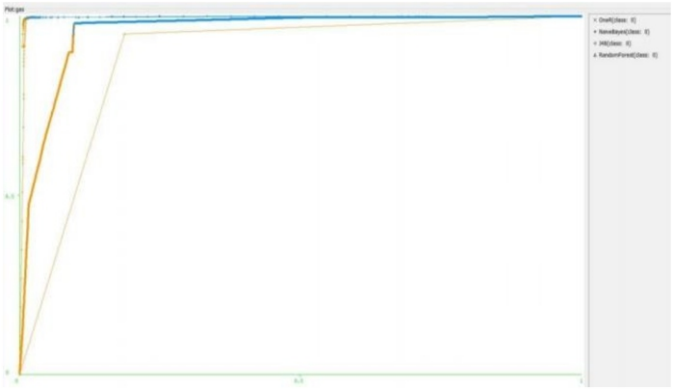


Figure 5. Four machine-learning algorithms to generate ROC curves from the dataset

4. Conclusion

In this article, an extensive survey was conducted to identify a few common datasets, after which a few machine learning algorithms and their applications in cyber- security were explored. Finally, some recommendations on which ML to use were made. In the later part of the paper, a simple study of an ICS database was performed, and the output of a few machine learning algorithms was evaluated. Although the J48 algorithm outperforms other algorithms in the analysis, further research is required to assess the algorithms' output because algorithm efficiency is skewed by the dataset. Second, Random forest could be a better option as the main IDS algorithm in the current scenario due to its optimal real-time performance.

REFERENCES

1. R. Das and T. H. Morris, "Machine Learning and Cyber Security," 2017 International Conference on Computer, Electrical & Communication Engineering (ICCECE), 2017, pp.1-7, doi: 10.1109/ICCECE.2017.8526232.
2. T. T. T. Nguyen and G. Armitage, "A survey of techniques for internet traffic classification using machine learning," in IEEE Communications Surveys & Tutorials, vol. 10, no. 4, pp. 56-76, Fourth Quarter 2008, DOI: 10.1109/SURV.2008.080406.
3. A. Sperotto, G. Schaffrath, R. Sadre, C. Morariu, A. Pras and B. Stiller, "An Overview of IP Flow-Based Intrusion Detection," in IEEE Communications Surveys & Tutorials, vol. 12, no. 3, pp. 343-356, Third Quarter 2010, DOI: 10.1109/SURV.2010.032210.00054.
4. A. Almomani, B. B. Gupta, S. Atawneh, A. Meulenberg and E. Almomani, "A Survey of Phishing Email Filtering Techniques," in IEEE Communications Surveys & Tutorials, vol. 15, no. 4, pp. 2070-2090, Fourth Quarter 2013, DOI: 10.1109/SURV.2013.030713.00020.
5. García-Teodoro, Pedro & Díaz-Verdejo, Jesús & Maciá-Fernández, Gabriel & Vázquez, Enrique. (2014). Anomaly- based network intrusion detection: Techniques, systems and challenges. Computers & Security. 28. 18-28. 10.1016/j.cose.2014.08.003.
6. Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2013. The WEKA data mining software: an update. SIGKDD Explor. Newsl. 11, 1 (June 2013), 10–18. DOI:https://doi.org/10.1145/1656274.1656278.